Sublanguages and Registers – A Note On Terminology *

Jussi Karlgren
Natural Language Processing Group, SICS
Box 1264, S-164 28 KISTA, Sweden
jussi@sics.se

October 25, 2005

Abstract

The term *sublanguage* from mathematical linguistics confuses interaction researchers and leads them to believe that implementing natural language interfaces is easier than it is. The term *register* from sociolinguistics is proposed instead.

Sublanguages and Interaction

Consider the following two observations:

1. The linguistic technology of today provides us with a patchy coverage of human verbal language at best.
2. Numerous studies show that people use specific languages to perform specific tasks.

It will be difficult, if not impossible, to construct a general natural language interface. The language the interface handles will be less general than the interaction researcher or language engineer would wish for. So, when using natural language interfaces, users will have to learn a specific language to perform tasks in. In general, natural language interfaces are constructed for specific tasks or domains considered appropriate for verbal interaction.

*Submission to *Interacting with Computers*

Given the two observations above, interaction researchers tend to discuss *sublanguages* (Rich 1985, Dahlbäck & Jönsson 1989, Véronis 1991, Diaper 1988) . A typical question can be posed as: "How can an appropriate *sublanguage* be chosen so that users can learn it efficiently?" (Rich 1985), or a research objective stated as " ... determining the *smallest complete sublanguage*" (Véronis 1991). The hope – although it seldom is explicitly put thus – is that a sublanguage will both be easier to implement for the language engineer, and will, if properly chosen, not hinder user interaction. This is of course a pipe dream, partly caused by the term itself. Using *sublanguage* in interface design obfuscates the issue.

The notion of a sublanguage has been given a precise mathematical definition by Zellig Harris (1968) as a subset of a well defined set. For unspoilt readers outside the interaction domain it will instil a vague mathematical uneasiness when used in this context. Using *sublanguage* implies that the specific task-oriented languages people use in specific situations are subsets of some *superlanguage*. Or, in other terms, that they are extracted from some natural language with superindividual qualities.

Superlanguages are not evident in naturally occuring usage of human language. The languages people use for specific tasks are not subsets of other language they use[1], nor extracted from the language they use in other situations In fact, they construct languages to suit the needs of a specific situation, even when this may mean using language they never have used before or never will again. An example, taken from a study of natural language interface users at IBM Nordic Laboratories (Karlgren 1992):

```
what was costed by consultants
```

"Costed" can hardly be assumed to be in most users' language to begin with. This is just one example of how overly simple the analysis of interaction languages as sublanguages of users' language.

Registers

In studies of human language in use and in sociolinguistic literature the mathematically less presumptious term *register* is used. A precise discussion of the term, which first apparently was used in the fifties by Reid, is carried out by M. A. K.

[1] A similar point has been made by Dan Diaper (1988) to motivate the study of sublanguages as languages in themselves rather than subsets.

Halliday (1978). Register is defined as a variety of language according to use, in contrast with varieties according to speaker or geographical location. In Halliday's words: "A register is what you are speaking, determined by what you are doing, and expressing diversity of social process. ... The register concept provides a means of investigating the linguistic foundations of everyday social interaction."

If the typical question above is reformulated as "How can an appropriate *register* be chosen so that users can learn it efficiently?" it will not raise unfounded hopes on the part of the language engineer or interface designer.

References

[1] Nils Dahlbäck and Arne Jönsson, "Empirical Studies of Discourse Representations for Natural Language Interfaces", *Proceedings of Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, 1989

[2] Dan Diaper, "Natural Language Communication with Computers: Theory, Needs and Practice." in *KBS in Government* P. Duffin (ed.), pp 19-44. Blenheim Online, 1988

[3] M. A. K. Halliday, *Language as social semiotic*, Edward Arnold Ltd., London, 1978

[4] Zellig Harris, *Mathematical Structures of Language*, John Wiley & Sons, New York, 1968

[5] Jussi Karlgren, *The Interaction of Discourse Modality and User Expectations in Human-Computer Dialog*, Licentiate Thesis at the Department of Computer and Systems Sciences, University of Stockholm, 1992

[6] Elaine Rich, "Natural Language Understanding: How Natural Can It Be?" , in *Proceedings of The Second Conference on Artificial Intelligence Applications*, 1985

[7] Jean Véronis, "Error in natural language dialogue between man and machine", *International Journal of Man-Machine Studies* 35, 1991, pp 187-217